AD NUMBER

AD460826

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; FEB 1965. Other requests shall be referred to Chemical Research and Develoment Laboratories, Edgewood Arsenal, MD 21010.

AUTHORITY

USAEA Notice, 19 Nov 1969

AD

CATALOGED BY: DDC

AS AD No. 460826

CRDL TECHNICAL MEMORANDUM 7-2

RAPID STRUCTURE SEARCHES VIA PERMUTED CHEMICAL LINE-NOTATIONS (III): A COMPUTER-PRODUCED INDEX

by

Mr. Charles E. Granito
Dr. John E. Schultz
1st Lt Gerald W. Gibson
Mr. Alan Gelberg
Mr. R. J. Williams
Dr. Edward A. Metcalf

February 1965

460826

US Army Edgewood Arsenal
CHEMICAL RESEARCH AND DEVELOPMENT LABORATORIES
Edgewood Arsenal, Maryland 21010

CRDL TECHNICAL MEMORANDUM 7-2

RAPID STRUCTURE SEARCHES VIA PERMUTED CHEMICAL LINE-NOTATIONS (III): A COMPUTER-PRODUCED INDEX

by

Mr. Charles E. Granito
Dr. John E. Schultz
1st Lt Gerald W. Gibson
Mr. Alan Gelberg
Mr. R. J. Williams
Dr. Edward A. Metcalf

INDUSTRIAL LIAISON OFFICE
OFFICE OF THE TECHNICAL DIRECTOR

FEBRUARY 1965

US Army Edgewood Arsenal
CHEMICAL RESEARCH AND DEVELOPMENT LABORATORIES
Edgewood Arsenal, Maryland 21010

FOREWORD

The work described in this memorandum was authorized under Task 1C522301A06001, Basic Agent Investigations (U). This phase of the permuted line-notation approach was started in June 1963 and completed in October 1964.

Notices

This memorandum is issued for temporary or limited use only, and it may be superseded.

Reproduction of this memorandum in whole or in part is prohibited except with permission of US Army Edgewood Arsenal Chemical Research and Development Laboratories; however, DDC is authorized to reproduce the document for US Government purposes.

Disclaimer

The findings in the memorandum are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Disposition

When this document has served its purpose, DESTROY it. DO NOT return the document to US Army Edgewood Arsenal Chemical Research and Development Laboratories.

RAPID STRUCTURE SEARCHES VIA PERMUTED CHEMICAL LINE-NOTATIONS (III): A COMPUTER-PRODUCED INDEX

by

Charles E. Granito[a], John E. Schultz[b],
Gerald W. Gibson, Alan Gelberg[c],
R. J. Williams[d], and E. A. Metcalf

INTRODUCTION

The previous papers in this series[1,2] have discussed the concept of an index of permuted Wiswesser chemical line-notations, the significance of a QUICK-SCAN area, and simple methods for preparing this type of index for a small index file of compounds (up to ca. 5000). It has been pointed out that the preparation of an index for a large number of compounds would require the use of a computer. This is the subject of this paper.

The project was started in 1963. At that time, a Univac File Computer (Model II) was readily available and, therefore, used for preparing the index. After the program was written and satisfactorily tested on a trial deck of 1000 cards, approximately 55,000 Wiswesser chemical line-notations, on file in this office, were indexed. A program which achieves this same result has been written for an IBM 1401 at the T. R. Evans Research Center of the Diamond Alkali Company. Both programs are available for potential users.

a - To whom all inquiries should be addressed.
b - Present Address: Westminster College, Fulton, Missouri.
c - Present Address: Diamond Alkali Company, Painesville, Ohio.
d - Data Processing Division, Management Science and Data Systems Office, Edgewood Arsenal, Maryland

The discussion of a general program to accomplish this permutation will be divided into four categories:

1. Computer preparation of the index.
2. Cost of preparing an index.
3. The index.
4. Uses of the index.

1. COMPUTER PREPARATION OF THE INDEX.

The input is a single punch-card per compound, containing an accession number, a two column screen, and a Wiswesser chemical line-notation. The program is designed to effect the permutation of the line-notation as each card is read onto magnetic tape, i.e., the operations required to select pertinent symbols, generate the scan area, and permute the notation are accomplished and the results stored prior to acceptance of the next line-notation. The path followed by a typical line notation card in these operations will be discussed rather than giving the step-by-step details of the less understandable directions and flow charts of the programmer.

For the purpose of this discussion, it is convenient to think of the input information as occupying one row of a compartment in the core memory (Memory I, figure 1). For easy visualization, the memory row which can hold 120 characters, is further divided into four areas: A, B, C, and D (figure 2). Each area is separated by a blank space to make the final print-out more readable, and the following arbitrary assignments are made:

a. Area A (8 spaces) for the accession number.
b. Area B (2 spaces) for a prefix to serve as a screen in the index.

c. Area C (11 spaces) for the QUICK-SCAN symbols to be generated by the computer.

d. Area D (96 spaces) for the line-notation.

Areas A, B, and D contain information read directly from the card. At the start, area C is empty; it will be filled by symbols, selected by the computer in its operations on a line-notation, for which an entry will be made in the index. The line-notation will be found in the last half of area D (spaces 73 through 120); initially, the first half of area D (spaces 25 through 73) is empty. Space 73, the center of area D, corresponds to the index column of the listing.

After all of the information on one card has been fed into the input section of Memory I, the computer is ready to start generating the information for the QUICK-SCAN area and the permutations of the line-notation. The notation is transferred to a reserve memory, Memory II (see figure 1). In this transfer and all subsequent transfers of the notation, the information in Areas A, B, and C is associated with the notation, but is not operated upon. Memory II holds the notation for a series of actions:

a. The notation is duplicated into Memory III, i.e., the information in spaces 73-120 of area D is duplicated.

b. In Memory III, the symbol in space 73, the indexing symbol, is compared to an exclusion list[2]. (A list of symbols used in the Wiswesser chemical line-notations that would not be useful as indexing terms.) If the symbol (space 73) is not

found on the list, the entire block of information is considered to be an index entry and is returned to Memory I for storage. Simultaneously, the index symbol (space 73) is duplicated and sent to Memory IV where it is held until all entries for this notation have been prepared. <u>The initial symbol of a notation is always an index entry</u>.

c. As the first block is accepted by Memory I a signal is sent to Memory II and again the notation is transferred to Memory III. However, this time the transfer involves a position change (one space left), so that the line-notation now occupies the 72-119 block, and the new symbol in space 73 is compared to the exclusion list. If the symbol is acceptable, the course of action is the same as for the first block of information. If the symbol is not a desirable indexing term, the information in Memory III is erased and Memory II is alerted by signal to continue the process.

d. As soon as Memory III compares two successive spaces containing no symbols (indicates end of line-notation) or after 120 (last position) this part of the operation ceases.

After each symbol in a line-notation has been considered and the appropriate operations have taken place on each, the symbols stored in Memory IV are returned to I and the SCAN SYMBOLS are entered in area C for each entry

that has been returned from the operations in Memory III. Prior to the acceptance of the second card, all information in Memory I is transferred to magnetic tape for storage and future sorting. The magnetic tape contains all the information generated from each line-notation; if printed-out at this stage, a listing as shown in figure 3, would be obtained. The entire operation from card to storage of the permuted line-notations on tape takes less than a second per card. The methodology described can be applied to line-notations containing a maximum of 48 columns* on a punch-card. In our files, approximately 85% of the compounds require twenty columns or less. In some cases, it is desirable to include in the notation field certain non-notational information, e.g., 95% pure, dimer, etc. To accomplish this the notation proper is followed by two spaces and the remaining area (up to space 120) is used for any desired information. Information preceded by two spaces will be printed out with the various permutations.

In order to create an index, all of the taped permutations must now be sorted alpha-numerically. This can be accomplished with an off-line tape sorter which requires no computer time. The tape of sorted notations is used to print out the index. The index column may be indicated either with a line (specially printed paper) or with a printer-produced mark referencing the index column at the top and bottom of each page. The example of a print-out (figure 4) shows a page of permuted notations for some of the sulfur compounds appearing in the Pesticide Index (2nd Edition). The address gives the page where the drawn structure and data may be found.

*This is an arbitrary assignment, any length may be chosen. For the Honeywell 400 we extended it to 60 columns and found that this covered greater than 99.9% of our file.

2. COST OF PREPARING AN INDEX.

The Univac File Computer (Model II) used for preparing the first large index is a relatively slow machine with a low tape density (92 characters per inch as compared to 536 for the Honeywell 400). It is, therefore, not comparable to most of the computers available in industry today which can perform this operation more economically. In order to make the dollar values more meaningful, the figures given will be based upon our experience with an IBM 1401 at the T. R. Evans Research Center and the Honeywell 400 which has replaced the Univac at the Data Processing Center of Edgewood.

The IBM 1401 was used for permuting 5991 line-notations. The input time was approximately 1¼ hours (80 cards per minute). For the 5991 compounds, a total of 33,080 entries were generated (5.5 entries/compound). These entries were stored on one-third of a reel of magnetic tape. In other words, one standard reel, (2400 ft.) could hold all of the entries for the indexing of about 18,000 compounds, or 108,000 entries (assuming 6 entries/compound). The alphanumeric sorting time required for the 33,080 entries was 30 minutes. The printing time (650 lines/minute) was 55 minutes. Therefore, the total machine time (input to hard copy) was just under 3 hours. At a rental of $52/hr. and allowing $15 for paper, the cost of preparing an index of the permuted line-notations of about 6,000 compounds, excluding labor and programming costs, would be approximately $165, i.e., 2.75 cents/compound.

Experience gained with the Honeywell 400 computer at Edgewood Arsenal leads us to predict that the 350,000 entries generated from 55,000 compounds (6.3 entries/compound) could be stored on 3 to 4 tapes whereas 21 tapes were required for the Univac II. Input requires about three hours, sorting time

4 hours, and printing time (900 lines/minute) 5 to 6 hours. Once again using a reasonable rental of $52/hr., the 55,000 compounds (350,000 entries) could be indexed for approximately $850, including $200 for paper. This averages out to about 1½ cents per compound.

In addition to actual operation costs, the cost of the program for the computer must be considered. This is a one time job for a given computer. The program written for the Univac took 100 hours (including debugging time). Since programming usually costs about $10/hour, $1000 would be a good estimate of the cost for writing the first program. However, the program subsequently written for the IBM 1401 required only 24 hrs. or $240. Since both programs are available on request, the cost for a program on a different model should be comparable to that required for the 1401. In other words, for about $1100 one could generate an index of permuted line-notations for 55,000 compounds.

Generation of 55,000 line-notations from chemical structures would cost about $10,000 (or 18 cents/notation). The rates used to arrive at this figure are summarized in figure 5. It should be remembered that notations, once prepared, will serve as input not only for the first index but subsequent ones as well.

3. THE INDEX

For 55,000 line-notations, an index of 7173 pages was obtained. There was a maximum of 49 entries per page, i.e., over 350,000 entries were generated. These pages were divided into 24 volumes. For ease of handling, each volume was cut to 11 x 14 inches and bound with hard covers. The back of

each volume was labeled in the same manner as an encyclopedia. After binding, the books occupied 40 inches of shelf space.

4. USE OF THE INDEX

In order to test the usefulness of the index some 25 structure searches were carried out. Each of these searches had been made previously using molecular formulae, a fragmentation code[3], and tabulated listings of line-notations in alpha-numeric order. In each case, the index was found to be significantly faster and a greater number of compounds meeting the search criteria were found. Some searches which had required a full day were completed in a matter of minutes using the index.

The index is used in the same manner as a dictionary. Both specific and general searches including analogs and homologs may be run at one's desk. There is no further need to use any mechanical equipment in locating desired structures.

a. Specific Look-up

For a specific structure, one prepares the line-notation and looks it up in the index. This process has proven to be faster than writing out a molecular formula and locating it in a molecular formula file. Each notation represents but one compound whereas each formula may represent many compounds. Since the notation is unique and unambiguous, it can be found in a specific location in the index. An average structure requires about 15 seconds to encode and a notation can be found in the index as fast as a word can be located in a dictionary.

b. General Searches

The procedure for running general searches is nearly as simple as the specific look-up. One decides which symbols meet the requirements of the request and either locates these in the index or assigns a clerical worker to the job. The frequency of occurrence for index symbols would determine the starting place in the index. Table 1 presents the frequency of index symbols for our first index. This table could be extended to include a more detailed break-down (beyond the first index symbol), if desired. It is not necessary to know what the line-notations represent to find entries in the index, just as it is not necessary to know the meaning of a word to find it in a dictionary.

With an index of permuted line-notations the individual in charge of chemical structure retrieval could rapidly answer such questions as:

(1) How many quinoline derivatives are on file?

(2) Which quinolines contain a nitro group as well as a hydroxyl group?

(3) How many aliphatic alkynes have been investigated and which ones contain chlorine?

All of these questions could be answered routinely. They would be handled as follows:

(1) Open the T volume of the index to the T66 BNJ section (a matter of seconds).

(2) Visually check the QUICK-SCAN area of the quinoline section, checking off those for which a NW

TABLE 1

FREQUENCY* OF OCCURRENCE FOR INDEX SYMBOLS (PERMUTED INDEX #1)

SYMBOL	RANK	TOTAL	SYMBOL	RANK	TOTAL
-	27	206	I	22	1966
1	17	4381	J	33	11
2	21	2132	K	20	3450
3	26	278	L	13	8421
4	25	647	M	8	20785
5	32	93	N	1	47201
6	29	151	O	2	38320
7	30	114	P	18	4029
8	28	174	Q	5	27513
9	31	96	R	4	34604
A	23	1731	S	9	19547
B	24	1361	T	6	24401
C	16	4565	U	10	17003
			V	3	35617
E	19	3624	W	12	8941
F	15	5430			
G	7	23524			
H	14	7502	Z	11	12453

*These frequencies are for <u>indexed</u> symbols, and do not include those excluded by the program (e.g., the T count is for T's which initiate ring systems, not T's indicating ring saturation).

and Q appear (for the nitro- and hydroxyl-groups, respectively).

(3) Open the index to the UU section, count the compounds not containing an L (carbocyclic), T (heterocyclic), or R (benzenoid) in the QUICK-SCAN area. Check off those which do have a G in the QUICK-SCAN area and list them.

A clerk with no chemical background could find the answers to each of the questions given above if provided with a list of appropriate search symbols. If a large number of searches were to be run every day this would be the most economical approach. The frequency of searches may well determine the need to perform computer searches. Since the permuted notations are already on magnetic tape, a search program can be written and performed.

When a large number of compounds are found that meet the search requirements, it has been found useful to make a Xerox reproduction of the index pages instead of writing down the numbers for each compound. This avoids transcription errors which mount up quickly for long lists of numbers.

It is realized that the forementioned questions are not complex. However, these are the types of questions most frequently asked. With the index, such questions take only minutes to answer. Any increase in the numbers of parameters or specificity shortens the search time required by reducing the search to a smaller section of the index.

c. Limitations

The only questions the index is not designed to handle efficiently are those involving the relative positions of every atom within a molecule, e.g.,

which compounds in the file contain a nitrogen atom three atoms from any oxygen atom and two carbons removed from a sulfur atom? In Wiswesser line-notations, symbols generally represent groups of atoms rather than an individual atom. This prevents the use of the index for such searches. However, it is not a limitation of the notation, for a computer could be programmed to use the notation for such searches[4].

d. Who Needs to Know the Notation

Only one or two chemists within an organization need to know the Wiswesser line-notation to make the index a useful means of retrieving chemical structures. The chemist or administrator can request information by structure(s) and receive answers in the same form. After compound numbers are found in the index, structure cards can be pulled, reproduced, and sent to the requester. The notation serves only as a means by which compounds are located.

e. Potential

As demonstrated, the index is a very powerful tool for keeping track of a file of chemical compounds. However, it has an even greater potential. For example, it could expedite procurement procedures. An organization which routinely purchases large numbers of materials from commercial suppliers could encode the compounds listed in available catalogues and include the source and price of each item. Preparation of an index would then permit rapid determination of compound availability, the companies offering an item, and their listed prices. All of this information would be found in the same section of the index, under the specific notation.

A second possibility is the generation of a functional group index for all chemical structures appearing in a catalog, textbook, journal, or secondary source publication.

Another possibility that is now being explored is the incorporation of biological data into the index. The inclusion of such data would permit investigation of structure-activity relationships. A glance at a given section would reveal how many compounds contained a given ring structure(s), functional group(s), or combinations thereof; as well as the type and level of activity exhibited. This could be a powerful aid to a Director of Research or his assistants. Any type of data could be included; the possibilities are unlimited. The data can be ordered by any desirable parameter with the linearized structures associated with it for comparison.

f. Updating the Index

Updating the index does not present a problem. Since the program is available, supplements, which will include compounds received after the major index was generated, can be prepared at suitable intervals. When the supplements become too numerous for easy searching, the tapes used for each index can be blended and used to create a new master index. The frequency of updating would depend upon the growth rate of the file.

SUMMARY

The preparation of a computer-produced index of permuted Wiswesser chemical line-notations is described. The uses and limitations of this powerful and economical retrieval tool are discussed. The utility of such an index may be markedly increased by the inclusion of biological, source, cost, etc., data.

REFERENCES

1. P. F. Sorter, C. E. Granito, J. C. Gilmer, A. Gelberg, and E. A. Metcalf, J.Chem.Doc. 4, 56 (1964).
2. C. E. Granito, A. Gelberg, J. E. Schultz, G. W. Gibson, and E. A. Metcalf, J.Chem.Doc. 5, 52 (1965).
3. A. Gelberg, W. Nelson, G. S. Yee, and E. A. Metcalf, J.Chem. Doc. 2, 7 (1962).
4. F. A. Landee, ACS Abstract of Papers, Div of Chem Lit, 3F7, 147th Meeting, April 6 - 9 (1964), Phila., Pa.

ACKNOWLEDGMENTS

The authors wish to thank Dr. D. H. Frear for permission to use the 2nd Edition of the Pesticide Index as a source for compounds to be coded and used as examples (figure 4).

The efforts and excellent cooperation of Mr. W. Nugent and Mr. W. Matthews, Diamond Alkali Company Data Center, for the IBM 1401 programming.

The support of the Edgewood Arsenal Data Processing Center for preparing the Univac program.

FIGURE 1

INFORMATION PATH FROM CARD TO TAPE

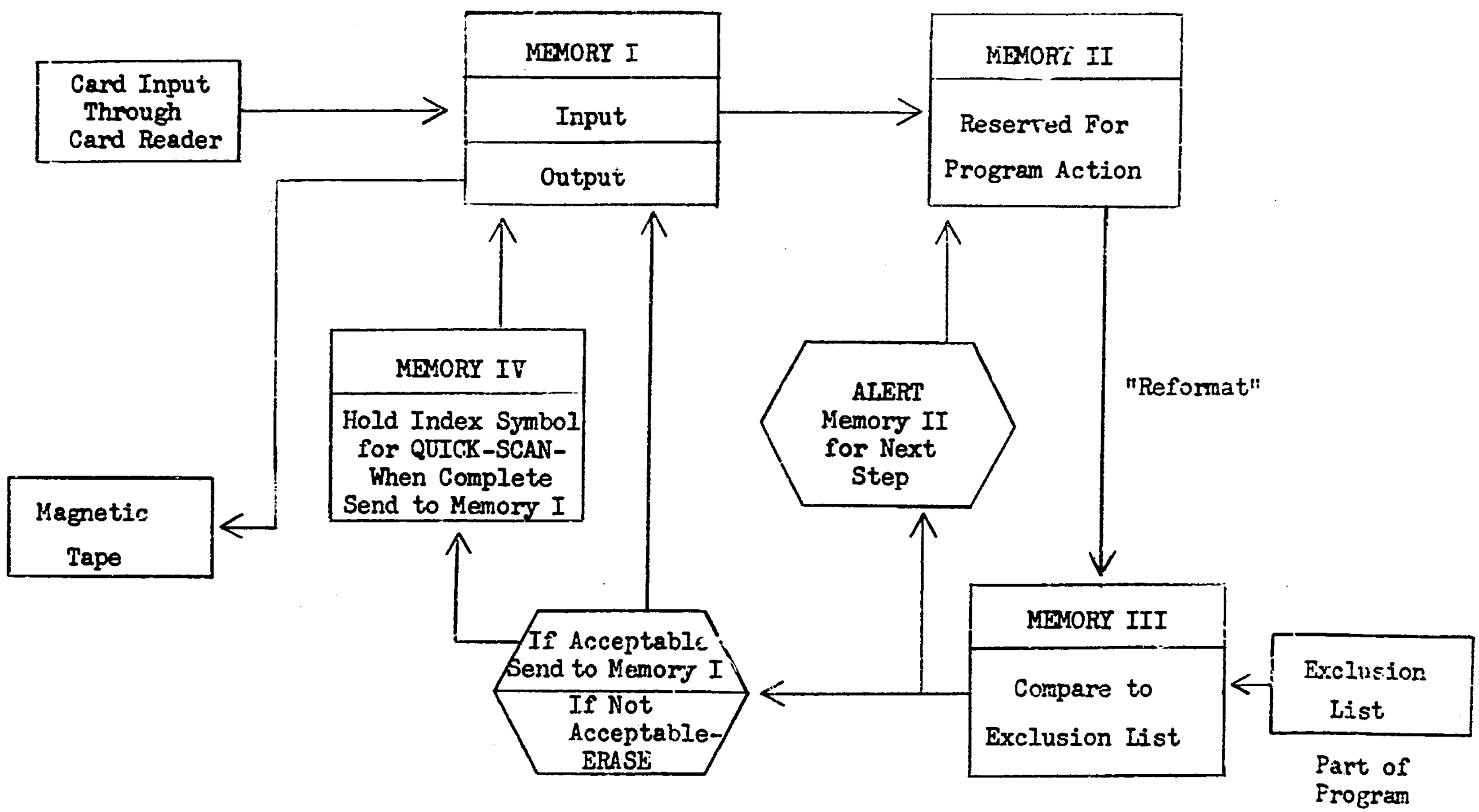

FIGURE 2

REPRESENTATION OF ONE SECTION OF INPUT-OUTPUT MEMORY IN COMPUTER

MEMORY I

Area	A		B		C		D	
Contents	ACCESSION NUMBER		PREFIX		QUICK-SCAN			NOTATION
Spaces	1-8	9	10-11	12	13-23	24	25-120	

73

FIGURE 3

EXAMPLE OF MAGNETIC TAPE PRINT-OUT PRIOR TO SORTING

```
                                  INDEX COLUMN
                                        ↓
00001008 C1 TNQ                         T6NJ BQ
00001008 C1 TNQ                      T6NJ BQ
00001008 C1 TNQ                   T6NJ BQ
00024101  6 QVG                         QV2G
00024101  6 QVG                       QV2G
00024101  6 QVG                    QV2G
```

FIGURE 4

EXAMPLE OF PRINT-OUT

PAGE OF PERMUTED NOTATIONS FOR SOME SULFUR COMPOUNDS APPEARING IN PESTICIDE INDEX (2ND EDITION)

COMPOUND NUMBER*	PREFIX	PERMUTED Symbols	NOTATION	Index Column
P12 130A	38	HGNRSWR	2-HG-NR&	SWR D
P12 152B	37	GSWRG	GJ	SWR DG
P12 211A	37	GRSWRG	G 3 DR L	SWR DG
P12 112C	38	ZSWROPOO	Z	SWR DOPO&/O1 2
P12 098E	38	ZSWROPOO	Z	SWR DOPO&/O2 2
P12 112D	38	ZSWROPSO	Z	SWR DOPS&/O1 2
P12 130F	38	MSWROPSO	2M	SWR DOPS&/O1 2
P12 149C	38	YMSWROPSO	1Y&M	SWR DOPS&/O1 2
P12 098F	38	ZSWROPSO	Z	SWR DOPS&/O2 2
P12 052C	37	ZSWRG	Z	SWR XG
P12 112E	B5	TNNRSWZNO	T5NNJ AR D	SWZ& C DNO E
P12 094A	8	OPSSXVM	10 2PS&	SX&&VM1
P12 047B	C7	TVNVUSXGGG	T56 BVNV GUTJ C	SXGGG
P12 135A	G7	TVNVSXGGG	T56 BVNVJ C	SXGGG
P12 210D	G7	TVNVUSXGGYGG	T56 BVNV GUTJ C	SXGGYGG
P12 139A	X5	TOSSXVNVNUUQQ	T66 I65 O 1C AS	SXVN GVN KU MU A BT&TTJ E F1Q JQ
P12 093E	8	OPSSSY	20 2PS&S1	SY
P12 222B	8	OPOSSYVM	10 2PO&S2	SY&M1
P12 090B	38	OPSSYRCN	20 2PS&	SYR&CN
P12 194C	P3	TNSYSNUS	T C566 BN D	SYS HNJ EUS
P12 213C	C3	TNSYSNUS	T C566 BN D	SYS HNJ EUS
P12 037B	3	NYUSSYUM	1N1&S	SYUS&M 22
P12 128C	3	NYUSSYUSM	1N1&YUS&	SYUS&M 22
P12 095B	38	OPSORS	20 2PS&OR C D	S1
P12 130D	38	OPSORS	2OPS&1&OR D	S1
P12 130E	38	OPSMCRS	2OPS&M1&OR C D	S1
P12 167C	38	SPORS	SP1&1&OR C D	S1
P12 100A	C3	TNNNSMY	T6N CN ENJ B	S1 D- F/MY 2
P12 200B	C3	TNNNSM	T6N CN ENJ B	S1 D--F/M2 2
P12 167B	C3	TNNNSMYM	T6N CN ENJ B	S1 DMY FM1
P12 024A	C3	TNNNSMYM	T6N CN ENJ B	S1 DMY FM2
P12 095C	38	OPSORS	20 2PS&OR C D	S1 E

*Refers to page number appearing in Pesticide Index

FIGURE 5

TIME AND COST ESTIMATES FOR AN INDEX OF PERMUTED LINE-NOTATIONS FOR 55,000 COMPOUNDS

INPUT (Line-notations)

	TIME (days)	COST ($)
WRITING NOTATIONS	110 (500/day)	4400 ($40/day)
PROOFING NOTATIONS	110	4400
PREPARING PUNCH-CARDS	27.5 (2000/day)	440 ($16/day)
VERIFYING PUNCH-CARDS	27.5	440
	TOTAL	$9680
COST OF 55,000 CARDS at $0.00125/CARD		69
COST FOR MACHINE RENTAL (KEY-PUNCH & VERIFIER)		120
		$9869

INDEX

PROGRAM COST		$240
MACHINE RENTAL (13 Hrs, $52/Hr.)		676
PAPER		200
	TOTAL	$1116

TOTAL COST FOR INDEXING 55,000 STRUCTURES	$12,085

DISTRIBUTION LIST

No. of Copies	
2	Technical Library, U. S. Army Chemical Research and Development Laboratories, Edgewood Arsenal, Maryland
2	Directorate of Defensive Systems, CRDL
2	Directorate of Technical Services, CRDL
2	Directorate of Medical Research, CRDL
2	Directorate of Weapons Systems, CRDL
2	Directorate of Development Support, CRDL
1	Nuclear Defense Laboratory, Edgewood Arsenal, Maryland
4	U. S. Army Biological Laboratories, Fort Detrick, Frederick, Maryland
1	Commanding General, U. S. Army Edgewood Arsenal, ATTN: Data Processing Division, Edgewood Arsenal, Maryland
1	Chief, U. S. Army Technical Information, Army Research Office, Washington 25, D. C.
1	Commanding General, Walter Reed Army Institute of Research, Washington 25, D. C.
20	Defense Documentation Center, Cameron Station, Alexandria, Virginia 22314
1	Commanding Officer, CRDL
1	Technical Director, CRDL
1	Mail and File Record Center, CRDL, RECORD COPY
1	Technical Clearances Section, Technical Releases Branch, CRDL
1	Publications Section, Technical Releases Branch, CRDL
50	Industrial Liaison Office, Office of the Technical Director, CRDL

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R&D	
(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)	
1. ORIGINATING ACTIVITY (Corporate author) US Army Edgewood Arsenal Chemical Research and Development Laboratories, Edgewood Arsenal, Maryland-Industrial Liaison Office, Office/Technical Director	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED; 2b. GROUP N/A
3. REPORT TITLE RAPID STRUCTURE SEARCHES VIA PERMUTED CHEMICAL LINE-NOTATIONS(III): A COMPUTER-PRODUCED INDEX	
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) The work was started in June 1963 and completed in October 1964	
5. AUTHOR(S) (Last name, first name, initial) Granito, Charles E., Schultz, John E., Gibson, Gerald W., Gelberg, Alan, Williams, R. J., Metcalf, Edward A.	
6. REPORT DATE February 1965	7a. TOTAL NO. OF PAGES 23; 7b. NO. OF REFS 4
8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S) CRDL Technical Memorandum 7-2
b. PROJECT NO. c. Task 1C522301A06001 d.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) N/A
10. AVAILABILITY/LIMITATION NOTICES Qualified requesters may obtain copies of this report from Defense Documentation Center, Cameron Station, Alexandria, Virginia 22314	
11. SUPPLEMENTARY NOTES Basic Agent Investigations	12. SPONSORING MILITARY ACTIVITY N/A

13. ABSTRACT

This is the third paper in a series concerned with the use of the Wiswesser line-notation as the basis of a storage and retrieval system for organic structures. This line-notation is an unique and unambiguous method of representing chemical structures by a linear series of letters, numbers, ampersands, and hyphens. The uniqueness of the notation permits the use of an alpha-numerically arranged index of the permutations of the notation for structure searching. The present paper discusses the preparation of a computer-prepared index of permuted Wiswesser line-notations from the standpoint of the cost, time, and logic of the computer program. The entire cost of preparing such an index of 50,000 structures, i.e., writing and proofing the notations, punching and verifying the punch-cards, card and paper costs, rental, and program preparation, is somewhat less than $12,000. The uses, limitations, and methods of further increasing the utility of this powerful and economical retrieval tool by the inclusion of biological, source, cost, and other types of data are also presented.

14. KEYWORDS

	Search-screen	Cost
Wiswesser line-notation	Keywords	Time
Chemical line-notation	Information retrieval	Program
Structure search	Automatic data retrieval system	Data
Quick-scan	Computer-prepared index	Permuted
	Updating	Index

DD FORM 1 JAN 64 1473

UNCLASSIFIED
Security Classification